PROJECT NUMBER : 0008
PROJECT TITLE : Computer Applications Division
WRITTEN BY : John Palesis and John Blankinship
PERIOD COVERED : February, 1991

I. Neurocomputing

A. Objective: Develop programming techniques for use of the HNC ANZA Plus Neurocomputer and Neurosoftware. Investigate potential applications of neurocomputing to R&D problems, and apply where appropriate.

B. Results: Research continued into several difficulties associated with backpropagation learning of multilayer networks. As manifested in the cigarette analyticals vs. liking problem and other mapping problems, these difficulties included: (1) extremely long training times with no guarantee of convergence to a good set of weights, (2) inability to train to a satisfactory mean squared error level, (3) overfitting of training data and poor generalization to new input patterns, (4) sensitivity to the initial set of weights and other learning parameters, resulting in significantly different predictions and standard errors depending on the random number seed, smoothing factors, etc., and (5) anamalous learning behavior.

The "MBPN Tool" control program and the training methodology for backpropagation learning were significantly enhanced to address these difficulties. The principal enhancements include: (1) testing mean squared performance during training with respect to a separate test data set to prevent overfitting of training data, (2) a cyclic mechanism for testing learning parameter adjustments during training, (3) allowing recovery after unsuccessful parameter adjustments and transient increases in the mean squared error, (4) a "stagnation monitor" to decrease the learning rate after the mean squared error has flattened out, (5) dynamically limiting the maximum learning rate, and (6) using an expanded training set. As a result of these enhancements, backpropagation learning is significantly improved in several respects. These include (1) faster convergence to a reliable set of weights resulting in lower mean squared error levels while still avoiding overfitting of training data, and (2) less sensitivity to the initial set of weights and other learning parameters.

Development of the Menthol Cigarette Liking Analysis Model (MCLAM), a neural network-based system for the analysis and prediction of liking ratings, was concluded. Using the enhanced backpropagation control program, a new neural network was determined for MCLAM. This network has five hidden layer processing elements, and exhibits a standard error of prediction of 0.43 on a seven-point liking scale. Also, a "Predict" option was added to MCLAM which determines predicted liking ratings against all smoker groups for existing or hypothetical menthol test products. The system is being used by PED in studies of menthol test products.

Using the HNC neurosoftware for Learning Vector Quantization (LVQ), a neural network technique for pattern recognition and classification was developed. LVQ iteratively determines a set of equiprobable representative patterns for each class

2022201702

Source: https://www.industrydocuments.ucsf.edu/docs/ptjm0000

given a training set of examples. Combined with the use of the Parzen Window technique for computing a conditional class probability, LVQ exhibits near Bayesian performance. The technique was applied to the pattern classification of certain electrophysiological waveforms. Although nearly 100 percent classification accuracy was achieved with respect to the training patterns, classification accuracy was only around 70 percent for new waveform patterns. This was attributed in part to the considerable variability in waveform patterns within each class.

C. Plans: Write a research report on the Menthol Cigarette Liking Analysis Model.

II. Expert Systems Development

A. Objective: Develop an Expert System for Cigarette Design.

B. Results: We are presently in the process of incorporating the new features of version 8.1 of the Fortran cigarette model into CigDES.1. This process essentially involves making changes and additions to (1) the interface between the Fortran mathematical model and the Lisp/KEE qualitative model and (2) the forward and goal-directed reasoning components of CigDES.1.

C. Conclusions: The task of identifying the changes and additions from version 7.2 to version 8.1 of the Fortran model is greatly facilitated by the use of the UNIX System V Source Control System (SCCS). SCCS compares the two different versions and lists their differences.

D. Plans: When this updating process is terminated, we will release CigDES.1.

E. References: Palesis, J.A., Dwyer, R.W., Leister, D.L., and Kao, J.W., "Transforming Mathematical Product Evaluation Models Into Expert Systems for Product Design," Proceedings of the 3rd International Conference on Industrial & Engineering Applications of Artificial Intelligence and Expert Systems, pp. 404-415, 1990.

III. Machine Learning

A. Objective: Apply AI-based machine learning in scientific research.

B. Results: "Popcorn-nutty" Odor Study: As reported in the monthly report of November 1990, ID3, an example-based machine learning algorithm has been used successfully to identify the underlying chemical substructures of ACYL-PYRIDINES which cause the "popcorn-nutty" odor. In this study, the structures of the chemical compounds analyzed by ID3 were represented using a "positional" notation, that is, as a set of structural positions (the attributes) whose values were unit atoms. In view of the fact that converting chemical compound structures to this "positional" notation requires a great deal of effort on the part of the chemist, we wanted to explore the possibility of applying ID3 directly to a set of structures represented in the "smiles" notation. This notation is used by many chemical databases to store compounds

2022201703

Source: https://www.industrydocuments.ucsf.edu/docs/ptjm0000

represented as strings of characters. The same set of compounds analyzed in the November study mentioned above were converted to the "smiles" notation and then processed by ID3. Unfortunately, the results were not satisfactory due to the fact that the ID3 algorithm analyzes the impact of single characters (i.e. columns in the relational table forming the "training set") on the result ("popcorn-nutty" odor in this case) and does not consider groups of characters (substrings) which represent meaningful units in the chemical compound structures.

C. Conclusions: ID3 is not a proper algorithm for analyzing compounds represented with the "smiles" notation. A more promising way to discover the underlying patterns of compounds represented as strings is to apply pattern recognition techniques similar to those used in natural language processing.

D. Plans: We will explore other methods for analyzing chemical compounds represented as strings to discover chemical substructures which cause specific chemical reaction.

E. References: (1) J. Palesis. Artificial Intelligence Based Induction: A Case for the ID3 Learning Algorithm. Philip Morris R&D Technical Report, January, 1991.

2022201704

Source: https://www.industrydocuments.ucsf.edu/docs/ptjm0000